UC DAVIS

DEPARTMENT OF ECONOMICS

Working Paper Series

Reasoning about strategies and rational play in dynamic games

Giacomo Bonanno
UC Davis

November 18, 2011

Paper # 11-11

We discuss a number of conceptual issues that arise in attempting to capture, in dynamic games, the notion that there is "common understanding" among the players that they are all rational.

Department of Economics
One Shields Avenue
Davis, CA 95616
(530)752-0741

http://www.econ.ucdavis.edu/working_search.cfm

Reasoning about strategies and rational play in dynamic games

Giacomo Bonanno*
Department of Economics,
University of California,
Davis, CA 95616-8578, USA
gfbonanno@ucdavis.edu

November 2011

Abstract

We discuss a number of conceptual issues that arise in attempting to capture, in dynamic games, the notion that there is "common understanding" among the players that they are all rational.

Keywords: Belief revision, common belief, counterfactual, dynamic game, model of a game, rationality.

1 Introduction

Game theory provides a formal language for the representation of interactive situations, that is, situations where several "entities" - called players - take actions that affect each other. The nature of the players varies according to the context in which the game theoretic language is invoked: in evolutionary

*This is the first draft of a chapter for a book on *Modeling strategic reasoning* edited by Johan van Benthem, Sujata Ghosh and RinekeVerbrugge.

biology (see, for example, [40]) players are non-thinking living organisms;[1] in computer science (see, for example, [39]) players are artificial agents; in behavioral game theory (see, for example, [24]) players are "ordinary" human beings, etc. Traditionally, however, game theory has focused on interaction among intelligent, sophisticated and rational individuals. For example Aumann describes game theory as follows:

> "Briefly put, game and economic theory are concerned with the interactive behavior of *Homo rationalis* - rational man. *Homo rationalis* is the species that always acts both purposefully and logically, has well-defined goals, is motivated solely by the desire to approach these goals as closely as possible, and has the calculating ability required to do so." ([3], p. 35.)

This chapter is concerned with the traditional interpretation of game theory and, in particular, with what is known as the epistemic foundation program, whose aim is to characterize, for any game, the behavior of rational and intelligent players who know the structure of the game and the preferences of their opponents and who recognize each other's rationality and reasoning abilities. The fundamental problem in this literature is to answer the following two questions: (1) under what circumstances can a player be said to be rational? and (2) what does 'mutual recognition' of rationality mean? While there seems to be agreement in the literature that 'mutual recognition' of rationality is to be interpreted as 'common belief' of rationality, the issue of what it means to say that a player is rational is not settled. Everybody agrees that the notion of rationality involves two ingredients: choice and be-

[1] Evolutionary game theory has been applied not only to the analysis of animal and insect behaviour but also to study the "most successful strategies" for tumor and cancer cells (see, for example, [30]).

liefs. However, the precise nature of their relationship involves subtle issues which will be discussed below, with a focus on dynamic games.

There is a bewildering collection of notions and results in the literature concerning the implications of rationality in dynamic games with perfect information: Aumann [4] proves that common *knowledge* of rationality implies the backward induction solution, Ben Porath [11] and Stalnaker [43] prove that common *belief/certainty* of rationality is *not* sufficient for backward induction, Samet [38] proves that what is needed for backward induction is common *hypothesis* of rationality, Feinberg [29] shows that common *confidence* of rationality logically contradicts the knowledge implied by the structure of the game, etc. The purpose of this chapter is not to review this literature[2] but to highlight some of the conceptual issues that have emerged.

In Section 2 we start with a brief exposition of one of the essential components of a definition of rationality, namely the concept of belief and we review the notions of model of a game and of rationality in the context of simultaneous games. We also discuss the role of counterfactuals in the analysis of simultaneous games. In the context of dynamic games there is a new issue that needs to be addressed, namely what it means to choose a strategy and what the proper interpretation of strategies is. This is addressed in Section 3 where we also discuss the subtle issues that arise when attempting to define rationality in dynamic games. In Section 4 we turn to the topic of belief revision in dynamic games and Section 5 concludes. The analysis is carried out entirely from a semantic perspective.[3]

[2]Surveys of this literature can be found in [9, 23, 28, 35].

[3]For a synactic analysis see [17, 25, 26, 27].

2 Belief, common belief and models of games

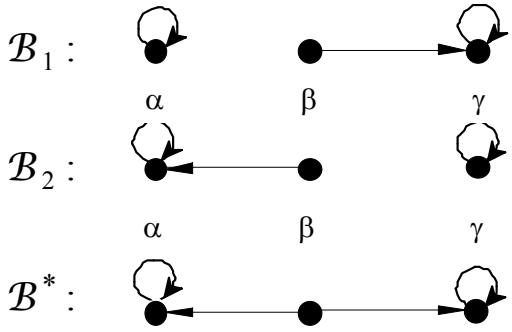
For simplicity we shall restrict attention to a qualitative notion of belief, thus avoiding the additional layer of complexity associated with probabilistic or graded beliefs. An *interactive belief structure* (or *multi-agent Kripke structure*) is a tuple $\langle N, \Omega, \{\mathcal{B}_i\}_{i\in N}\rangle$ where N is a finite set of *players*, Ω is a set of *states* and, for every player i, $\mathcal{B}_i$ is a binary relation on Ω representing *doxastic accessibility*: $\mathcal{B}_i(\omega) \stackrel{def}{=} \{\omega' \in \Omega : \omega \mathcal{B}_i \omega'\}$ is the set of states that are compatible with player i's beliefs at state ω.[4] We assume that each $\mathcal{B}_i$ is serial ($\mathcal{B}_i(\omega) \neq \varnothing,\ \forall \omega \in \Omega$), transitive (if $\omega' \in \mathcal{B}_i(\omega)$ then $\mathcal{B}_i(\omega') \subseteq \mathcal{B}_i(\omega)$) and euclidean (if $\omega' \in \mathcal{B}_i(\omega)$ then $\mathcal{B}_i(\omega) \subseteq \mathcal{B}_i(\omega')$). Seriality captures the notion of consistency of beliefs, while the last two properties correspond to the notions of positive and negative introspection of beliefs.[5] A subset E of Ω is called an *event*. Associated with the binary relation $\mathcal{B}_i$ is a *belief operator* on events $B_i : 2^\Omega \to 2^\Omega$ defined by $B_i E = \{\omega \in \Omega : \mathcal{B}_i(\omega) \subseteq E\}$. Thus $B_i E$ is the event that player i believes E. Figure 1a shows an interactive belief structure with two players, where each relation $\mathcal{B}_i$ is represented by arrows: $\omega' \in \mathcal{B}_i(\omega)$ if and only if there is an arrow from ω to ω'. In this structure we have, for example, that $B_1\{\gamma\} = \{\beta, \gamma\}$, that is, at both states β and γ Player 1 believes event $\{\gamma\}$. Let $\mathcal{B}^*$ be the transitive closure of $\bigcup_{i\in N} \mathcal{B}_i$ [6] and define the operator $B^* : 2^\Omega \to 2^\Omega$ by $B^* E = \{\omega \in \Omega : \mathcal{B}^*(\omega) \subseteq E\}$. B^* is called the *common belief* operator and when $\omega \in B^* E$ then at ω

[4] Thus $\mathcal{B}_i$ can also be viewed as a function from Ω into 2^Ω (the power set of Ω). Such functions are called *possibility correspondences* in the game-theoretic literature.

[5] For more details see the survey in [9].

[6] That is, $\omega' \in \mathcal{B}^*(\omega)$ if and only if there is a sequence $\langle \omega_1, ..., \omega_m\rangle$ in Ω and a sequence $\langle j_1, ..., j_{m-1}\rangle$ in N such that (1) $\omega_1 = \omega$, (2) $\omega_m = \omega'$ and (3) for all $k = 1, ..., m-1$, $\omega_{k+1} \in \mathcal{B}_{j_k}(\omega_k)$.

every player believes E and every player believes that every players believes E, and so on, *ad infinitum*. Figure 1a shows the relation $\mathcal{B}^*$ (the transitive closure of $\mathcal{B}_1 \cup \mathcal{B}_2$): in this case we have that, for example, $B^*\{\gamma\} = \{\gamma\}$ but $B_1B^*\{\gamma\} = \{\beta, \gamma\}$, that is, event $\{\gamma\}$ is commonly believed only at state γ but at state β Player 1 erroneously believes that it is common belief that $\{\gamma\}$ is the case.



			Player 2	
			l	*r*
Player 1	*t*		2 , 1	0 , 0
	b		1 , 2	1 , 2

(a) An interactive belief structure (b) A strategic-form game

Figure 1

When the relations $\mathcal{B}_i$ ($i \in N$) are also assumed to be reflexive ($\omega \in \mathcal{B}_i(\omega)$, $\forall \omega \in \Omega$), then they become equivalence relations and thus each $\mathcal{B}_i$ gives rise to a partition of Ω. In partitional models, beliefs are necessarily correct and one can speak of *knowledge* rather than belief. As Stalnaker [42] points out, it is methodologically preferable to carry out the analysis in terms of (possibly erroneous) beliefs and then - if desired - add further conditions that are sufficient to turn beliefs into knowledge. The reason why one should not start with the assumption of necessarily correct beliefs (that is, reflexivity of $\mathcal{B}_i$) is that such an assumption has strong intersubjective implications:

> "The assumption that Alice believes (with probability one) that Bert believes (with probability one) that the cat ate the canary

> tells us nothing about what Alice believes about the cat and the canary themselves. But if we assume instead that Alice knows that Bert knows that the cat ate the canary, it follows, not only that the cat in fact ate the canary, but that Alice knows it, and therefore believes it as well." ([42] p. 153.)

One can locally (that is, at a state ω) transform belief into knowledge by adding the hypothesis that at least one player has correct beliefs (for some $i \in N$, $\omega \in \mathcal{B}_i(\omega)$) and that there is common belief that nobody has erroneous beliefs (for all $\omega' \in \mathcal{B}^*(\omega)$ and for all $i \in N$, $\omega' \in \mathcal{B}_i(\omega')$). Adding such a hypothesis introduces strong forms of agreements among the players (see [22]) and is, in general, not realistic.

Interactive belief structures can be used to model particular contexts in which a game is played. Let us take, as a starting point, strategic-form games (also called normal-form games), where players make their choices simultaneously (an example is a sealed-bid auction). A *strategic-form game* is a tuple $\left\langle N, \{S_i, \succsim_i\}_{i \in N} \right\rangle$ where N is a set of *players* and, for every $i \in N$, S_i is a set of choices or *strategies* available to player i and $\succsim_i$ is i's preference relation over the set of *strategy profiles* $S = \underset{i \in N}{\times} S_i$.[7] We shall throughout focus on ordinal preferences (rather than cardinal preferences with associated expected utility comparisons) for two reasons: (1) since the

[7] A preference relation over a set S is a binary relation $\succsim$ on S which is complete or connected (for all $s, s' \in S$, either $s \succsim s'$ or $s' \succsim s$, or both) and transitive (for all $s, s', s'' \in S$, if $s \succsim s'$ and $s' \succsim s''$ then $s \succsim s''$). We write $s \succ s'$ as a short-hand for $s \succsim s'$ and $s' \not\succsim s$ and we write $s \sim s'$ as a short-hand for $s \succsim s'$ and $s' \succsim s$. The interpretation of $s \succsim_i s'$ is that player i considers s to be at least as good as s', while $s \succ_i s'$ means that player i prefers s to s' and $s \sim_i s'$ means that she is indifferent beween s and s'.

The interpretation is that there is a set Z of possible outcomes over which every player has a preference ralation. An outcome function $o : S \to Z$ associates an outcome with every startegy profile, so that the preference relation over Z induces a preference relation over S.

game is usually hypothesized to be common knowledge among the players, it seems far more realistic to assume that each player knows the ordinal rankings of her opponents rather than their full attitude to risk (represented by a cardinal utility function) and (2) our aim is to point out some general conceptual issues, which are independent of the notion of expected utility.

The definition of strategic-form game specifies the choices available to the players and what motivates those choices (their preferences over the possible outcomes); however, it leaves out an important factor in the determination of players' choices, namely what they believe about the other players. Adding a specification of the players' beliefs determines the context in which a particular game is played and this can be done with the help of an interactive belief structure. Fix a strategic-form game $G = \langle N, \{S_i, \succsim_i\}_{i \in N} \rangle$. A *model of* G is a tuple $\langle N, \Omega, \{\mathcal{B}_i, \}_{i \in N}, \{\sigma_i, \}_{i \in N} \rangle$, where $\langle N, \Omega, \{\mathcal{B}_i, \}_{i \in N} \rangle$ is an interactive belief structure and, for every $i \in N$, $\sigma_i : \Omega \to S_i$ is a function that assigns to each state ω a strategy $\sigma_i(\omega) \in S_i$ of player i. Let $\sigma(\omega) = (\sigma_i(\omega))_{i \in N}$ denote the strategy profile associated with state ω. The function $\sigma : \Omega \to S$ gives content to the players' beliefs. If $\omega \in \Omega$, $x \in S_i$ and $\sigma_i(\omega) = x$ then the interpretation is that at state ω player i "chooses" strategy x. The exact meaning of 'choosing' is not elaborated further in the literature: does it mean that player i *has actually played* x, or that she is *committed to playing* x, or that x is the *output of her deliberation process*? Whatever the answer, the assumption commonly made in the literature is that player i has correct beliefs about her chosen strategy, that is, she chooses strategy x if and only if she believes that her chosen strategy is x. This can be expressed formally as follows. For every $x \in S_i$, let $[\sigma_i = x]$ be the event that player i chooses

strategy x, that is, $[\sigma_i = x] = \{\omega \in \Omega : \sigma_i(\omega) = x\}$. Then the assumption is that

$$[\sigma_i = x] = B_i [\sigma_i = x] . \tag{1}$$

We will return to this assumption later on, in our discussion of dynamic games. Figure 1*b* shows a strategic-form game in the form of a table, where the preference relation $\succsim_i$ of player i is represented numerically by an ordinal *utility function* $u_i : S \to \mathbb{R}$, that is, a function satisfying the property that $u_i(s) \geq u_i(s')$ if and only if $s \succsim_i s'$. In each cell of the table the first number is the utility of Player 1 and the second number the utility of Player 2. A model of this game can be obtained by adding to the interactive belief frame of Figure 1*a* the following strategy assignments:

$$\begin{aligned} &\sigma_1(\alpha) = b, \quad \sigma_1(\beta) = \sigma_1(\gamma) = t \\ &\sigma_2(\alpha) = \sigma_1(\beta) = r, \quad \sigma_2(\gamma) = l. \end{aligned} \tag{2}$$

How can rationality be captured in a model? Consider the following - rather weak - definition of rationality: player i is rational at state $\hat{\omega}$ if - letting $\hat{s}_i = \sigma_i(\hat{\omega}) \in S_i$ - there is no other strategy $s_i \in S_i$ which player i believes to be better than $\hat{s}_i$. This can be stated formally as follows. First of all, for every state ω, denote by $\sigma_{-i}(\omega)$ the strategy profile of the players other than i, that is, $\sigma_{-i}(\omega) = (\sigma_1(\omega), ..., \sigma_{i-1}(\omega), \sigma_{i+1}(\omega), ..., \sigma_n(\omega))$ (where n is the cardinality of N). Then (recall that, by (1) - since $\sigma_i(\hat{\omega}) = \hat{s}_i$ - for all $\omega \in \mathcal{B}_i(\hat{\omega})$, $\sigma_i(\omega) = \hat{s}_i$):

$$\begin{array}{l}\text{Player } i \text{ is } \textit{rational at } \hat{\omega} \text{ if, } \forall s_i \in S_i, \text{ it is not the case} \\ \text{that, } \forall \omega \in \mathcal{B}_i(\hat{\omega}),\ \ u_i\left(s_i, \sigma_{-i}(\omega)\right) > u_i\left(\hat{s}_i, \sigma_{-i}(\omega)\right).\end{array} \tag{3}$$

Equivalently, let $[u_i(s_i) > u_i(\hat{s}_i)] = \{\omega \in \Omega : u_i\left(s_i, \sigma_{-i}(\omega)\right) > u_i\left(\hat{s}_i, \sigma_{-i}(\omega)\right)\}$. Then (recall that $\sigma_i(\hat{\omega}) = \hat{s}_i$)

$$\text{Player } i \text{ is } \textit{rational at } \hat{\omega} \text{ if, } \forall s_i \in S_i,\ \hat{\omega} \notin B_i\left[u(s_i) > u_i(\hat{s}_i)\right]. \tag{4}$$

For example, in the model of the strategic-form game of Figure 1*b* obtained by adding to the interactive belief structure of Figure 1*a* the strategy assignments given above in (2), we have that both players are rational at every state and thus there is common belief of rationality at every state. In particular, there is common belief of rationality at state β, even though the strategy profile actually chosen there is (t, r) (with payoffs $(0, 0)$) and each player would do strictly better with a different choice of strategy. Note also that, in this model, at every state it is common belief between the players that each player has correct beliefs,[8] although at state β neither player does in fact have correct beliefs.

It is well known that, in any model of any finite strategic-form game, a strategy profile $s = (s_i)_{i \in N}$ is compatible with common belief of rationality if and only if, for every player i, the strategy s_i survives the iterated deletion of strictly dominated strategies.[9]

[8] $\forall \omega \in \Omega$, $\forall \omega' \in \mathcal{B}^*(\omega)$, $\omega' \in \mathcal{B}_1(\omega')$ and $\omega' \in \mathcal{B}_2(\omega')$.

[9] That is, if at a state ω there is common belief of rationality then, for every player i, $\sigma_i(\omega)$ survives the iterated deletion of strictly dominated strategies. For a proof and more details see [17].

What is the conceptual content of the definition given in (4)? It is widely claimed that the notion of rationality involves the (implicit or explicit) use of counterfactual reasoning. For example Aumann writes:

> "[...] one really cannot discuss rationality, or indeed decision making, without substantive conditionals and counterfactuals. Making a decision means choosing among alternatives. Thus one must consider hypothetical situations - what would happen if one did something different from what one actually does. [...] In interactive decision making - games - you must consider what other people would do if you did something different from what you actually do." ([4], p. 15.)

Yet the structures used so far do not incorporate the tools needed for counterfactual reasoning. The definition of rationality given in (4) involves comparing the payoff of a strategy different from the one actually chosen with the payoff of the chosen strategy. Can this counterfactual be made explicit?

First we review the standard semantics for counterfactuals introduced by Stalnaker [41]. Given a set of states Ω and a set $\mathcal{E} \subseteq 2^{\Omega} \backslash \varnothing$ of events interpreted as admissible hypotheses, a *counterfactual selection function* is a function $f : \Omega \times \mathcal{E} \to \Omega$ that satisfies the following properties: $\forall \omega \in \Omega$, $\forall E, F \in \mathcal{E}$,

$$\begin{array}{ll} 1. & f(\omega, E) \in E, \\ 2. & \text{if } \omega \in E \text{ then } f(\omega, E) = \omega, \\ 3. & \text{if } f(\omega, E) \in F \text{ and } f(\omega, F) \in E \text{ then } f(\omega, E) = f(\omega, F). \end{array} \tag{5}$$

If $\omega' = f(\omega, E)$ then the interpretation is that ω' is the state closest (or most similar) to ω where hypothesis E is true.[10] Condition 1 is a consistency condition that says that the state closest to ω where E is true is indeed a state where E is true. Condition 2 says that if E is true at ω then the state most similar to ω where E is true is ω itself. Condition 3 says that, if the closest E-state to ω is in F and the closest F-state to ω is in E, then two states must coincide.

Given a hypothesis $E \in \mathcal{E}$ and an event $F \subseteq \Omega$, a counterfactual statement of the form "if E were the case then F would be the case", which we denote by $E \rightrightarrows F$, is considered to be true at state ω if and only if $f(\omega, E) \in F$, that is, if F is true in the closest world to ω where E is true. Correspondingly, one can define the operator $\rightrightarrows \ : \mathcal{E} \to 2^{\Omega}$ as follows:

$$E \rightrightarrows F = \{\omega \in \Omega : f(\omega, E) \in F\}. \tag{6}$$

Adding a counterfactual selection function to an interactive belief structure allows one to consider complex statements of the form "if E were the case then player i would believe F" (corresponding to the event $E \rightrightarrows B_i F$), or "player i believes that if E were the case then F would be the case" (corresponding to $B_i(E \rightrightarrows F)$), or "Player 1 believes that if E were the case then Player 2 would believe F" (corresponding to $B_1(E \rightrightarrows B_2 F)$), etc.

Now, returning to models of strategic-form games and the definition of rationality given in (4), the addition of a counterfactual selection function

[10] We chose the simpler version of the theory, due to Stalnaker, where $f(\omega, E)$ is a single element of Ω. The approach was later generalized by Lewis [34] by allowing $f(\omega, E)$ to be a subset of E. For a review of the general approach see [32]. For the conceptual points that we want to highlight there is no loss of generality in adopting Stalnaker's approach.

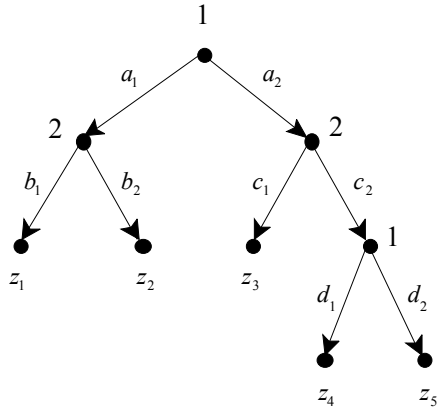
to a model allows one to compare player i's payoff at a state $\hat{\omega}$ where he has chosen strategy $\hat{s}_i$ with her payoff at the state closest to $\hat{\omega}$ where she chooses a strategy $s_i \neq \hat{s}_i$. Implicit in (4) is the assumption that in that counterfactual world player i's beliefs about her opponents' choices are the same as in $\hat{\omega}$. This is an assumption: it may be a sensible one to make (indeed Stalnaker [43, 44] argues that it would be conceptually wrong *not* to make this assumption) but nonetheless it may be worthwhile bringing it to light in a more complete analysis where counterfactuals are explicitly modeled. Within the context of strategic-form games, this is done in [14, 48], where counterfactuals are invoked explicitly in the definition of rationality.

3 Models of dynamic games

In dynamic (or extensive-form) games players make choices sequentially having some information about the moves previously made by their opponents. If information is partial, the game is said to have *imperfect information*, while the case of full information is referred to as *perfect information*. An example of the latter is shown in Figure 2*a* in the form of a tree (the players' preferences over outcomes have been omitted). Each node in the tree represents a history of prior moves and is labeled with the player whose turn it is to move. For example, at history a_2c_2 it is Player 1's turn to move (after his initial choice of a_2 followed by Player 2's choice of c_2) and he has to choose between two actions: d_1 and d_2. The terminal histories (the leaves of the tree, denoted by z_j, $j = 1, ..., 5$) represent the possible outcomes and each player i is assumed to have a preference relation $\succsim_i$ over the set of terminal

histories.

In their seminal book, von Neumann and Morgenstern [47] showed that a dynamic game can be reduced to a normal-form game by defining strategies as complete, contingent plans of action. In the case of perfect-information games a strategy for a player is a function that associates with every history assigned to that player one of the choices available there. For example, a possible strategy of Player 1 in the game of Figure 2a is (a_1, d_2). A profile of strategies (one for each player) determines a unique path from the null history (the root of the tree) to a terminal history (a leaf of the tree). Figure 2b shows the strategic-form corresponding to the extensive form of Figure 2a.



Player 1 \ Player 2	b_1c_1	b_1c_2	b_2c_1	b_2c_2
a_1d_1	z_1	z_1	z_2	z_2
a_1d_2	z_1	z_1	z_2	z_2
a_2d_1	z_3	z_4	z_3	z_4
a_2d_2	z_3	z_5	z_3	z_5

(a) A perfect-information game (b) The corresponding strategic form

Figure 2

How should a model of a dynamic game be constructed? One approach in the literature (see, for example, [4]) has been to consider models of the corresponding strategic-form (the type of models considered in Section 2). However, there are several conceptual issue that arise in this context. Recall

that the interpretation of $s_i = \sigma_i(\omega)$ suggested in Section 2 is that at state ω player i "chooses" strategy s_i. Now consider a model of the game of Figure 2a and a state ω where $\sigma_1(\omega) = (a_1, d_2)$. What does it mean to say that Player 1 "chooses" strategy (a_1, d_2)? The first part of the strategy, namely a_1, can be interpreted as the decision by Player 1 to play a_1, but the second part of the strategy, namely d_2, has no such interpretation: if Player 1 in fact plays a_1 then he knows that he will not have to make any further choices and thus it is not clear what it means to "choose" to play d_2 in a situation that is made impossible by his decision to play a_1. Thus it does not seem to make sense to interpret $\sigma_1(\omega) = (a_1, d_2)$ as 'at state ω Player 1 chooses (a_1, d_2)'. Perhaps the correct interpretation is in terms of a more complex sentence such as 'Player 1 chooses to play a_1 and if - contrary to this - he were to play b_1 and Player 2 were to follow with c_2 then Player 1 would play d_2'. Thus while in a simultaneous game the association of a strategy of player i to a state can be interpreted as a description of player i's behavior at that state, in the case of dynamic games this interpretation is no longer valid, since one would end up describing not only the actual behavior of player i but also his counterfactual behavior at a different state. Methodologically this is not a satisfactory choice: if it is necessary to specify what a player would do in situations that do not occur in the state under consideration, then one should model the counterfactual explicitly. But why should it be necessary to specify at state ω (where Player 1 is playing a_1) what he would do at the counterfactual history a_2c_2? Perhaps what matters is not so much what Player 1 would actually do there but what Player 2 believes that Player 1 would do: after all, Player 2 might not know that Player 1 has decided to

play a_1 and needs to consider what to do in the eventuality that Player 1 actually ends up playing a_2. So perhaps, the strategy of Player 1 is to be interpreted not as a description of Player 1's behavior but as a conjecture in the mind of Player 2 about what Player 1 would do. This interpretation of strategies has in fact been put forward in the literature for the case of mixed strategies (which we will not consider in this chapter, given our non-probabilistic approach).[11]

In order to clarify these issues it seems that, in the case of dynamic games, one should not adopt the models of Section 2 and instead consider a more general notion of model, where states are described in terms of players' actual behavior and any relevant counterfactual propositions are modeled explicitly. For simplicity we will focus on perfect-information games. Fix a dynamic game Γ with perfect information and consider the following candidate for a definition of a model of Γ: a model of Γ is a tuple $\left\langle N, \Omega, \{\mathcal{B}_i\}_{i \in N}, f, \zeta \right\rangle$ where $\left\langle N, \Omega, \{\mathcal{B}_i\}_{i \in N} \right\rangle$ is an interactive belief structure, $f : \Omega \times \mathcal{E} \to \Omega$ is a counterfactual selection function and $\zeta : \Omega \to Z$ is a function that associates with every state $\omega \in \Omega$ a terminal history (Z denotes the set of terminal histories in Γ).[12] Given a history h in the game, we denoted by $[h]$ the set of states where h is reached, that is, $[h] = \{\omega \in \Omega : h \text{ is a prefix of } \zeta(\omega)\}$. We take the set of admissible hypotheses $\mathcal{E}$ (the domain of $f(\omega, \cdot)$) to be the set of propositions of the form "history h is reached", that is, $\mathcal{E} = \{[h] : h \in H\}$ (where H is the set of histories in the game). We now discuss a number of issues that arise in such models.

[11] See, for example, [6] and the references given there in Footnote 7.

[12] Samet [38] was the first to propose models of perfect-information games where states are described not in terms of strategies but in terms of terminal histories.

In the models of Section 2 it was assumed that a player always knows his own strategy (see (1) above). Should a similar assumption be made within the context of dynamic games? That is, suppose that at state ω player i takes action a; should we assume that player i believes that she takes action a? For example, consider a model of the game of Figure 2a and two states, ω and ω' such that $\mathcal{B}_2(\omega) = \{\omega, \omega'\}$ and $\zeta(\omega) = a_1 b_1$. Then at state ω Player 2 takes action b_1. Should we require that Player 2 take action b_1 also at ω' (since $\omega' \in \mathcal{B}_2(\omega)$)? The answer is negative: Player 2 may be uncertain as to whether Player 1 will play a_1 or a_2 and plan to play herself b_1 in the former case and c_1 in the latter case. Thus it makes perfect sense to have $\zeta(\omega') = a_2 c_1$. If we want to rule out uncertainty by a player about her action at a decision history of hers, then we need to impose the following restriction:

$$\begin{array}{l} \text{If } h \text{ is a decision history of plater } i, \ a \text{ an action at } h \\ \text{and } ha \text{ a prefix of } \zeta(\omega) \text{ then, } \forall \omega' \in \mathcal{B}_i(\omega), \\ \text{if } h \text{ is a prefix of } \zeta(\omega') \text{ then } ha \text{ is a prefix of } \zeta(\omega'). \end{array} \tag{7}$$

The above definition can be stated more succinctly in terms of events. If E and F are two events, we denote by $E \to F$ the event $\neg E \cup F$. Thus $E \to F$ captures the material conditional. Recall that, given a history h in the game, $[h] = \{\omega \in \Omega : h \text{ is a prefix of } \zeta(\omega)\}$. Let H_i denote the set of decision histories of player i and $A(h)$ the set of choices available at h. Then (7) can be stated as follows:[13]

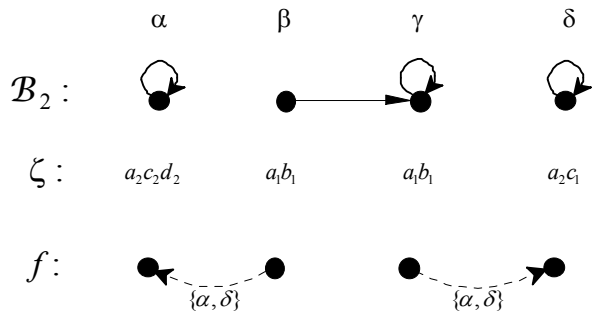
[13] Note that, if at state ω player i believes that history h will *not* be reached ($\forall \omega' \in \mathcal{B}_i(\omega)$, $\omega' \notin [h]$) then $\mathcal{B}_i(\omega) \subseteq \neg[h] \subseteq [h] \to [ha]$, so that $\omega \in B_i([h] \to [ha])$ and therefore (8) is satisfied even if $\omega \in [ha]$.

$$\begin{aligned} &\forall h \in H_i, \forall a \in A(h), \\ &[ha] \subseteq B_i([h] \rightarrow [ha]). \end{aligned} \tag{8}$$

In words: if at a state player i takes action a at her decision history h, then she believes that if h is reached then she takes action a.

A more subtle issue is whether we should require (perhaps as a condition of rationality) that a player have correct beliefs about what she would do in a situation that she believes will not arise. Consider, for example, the (part of a) model of the game of Figure 2a illustrated in Figure 3. The first line gives $\mathcal{B}_2$, the doxastic accessibility relation of Player 2, the second line the function ζ (which associates with every state a terminal history) and the third line is a partial illustration of the counterfactual selection function: in particular we have that $f(\beta, \{\alpha, \delta\}) = \alpha$ and $f(\gamma, \{\alpha, \delta\}) = \delta$. Note that the event that Player 1 plays a_2 is $[a_2] = \{\alpha, \delta\}$. Recall that $E \rightrightarrows F$ denotes the counterfactual conditional 'if E were the case then F would be the case'. Now, $[a_2] \rightrightarrows [a_2c_1] = \{\gamma, \delta\}$ and $[a_2] \rightrightarrows [a_2c_2] = \{\alpha, \beta\}$. Thus $\beta \in [a_2] \rightrightarrows [a_2c_2]$ and $\beta \in B_2([a_2] \rightrightarrows [a_2c_1])$.[14] That is, at state β it is actually the case that if Player 1 were to play a_2 then Player 2 would respond with c_2, but Player 2 erroneously believes that (if Player 1 were to play a_2) she would respond with c_1.

[14] Recall that the material conditional 'if E is the case then F is the case' is captured by the event $\neg E \cup F$, which we denote by $E \rightarrow F$. Then $[a_2] \rightarrow [a_2c_1] = \{\beta, \gamma, \delta\}$ and $[a_2] \rightarrow [a_2c_2] = \{\alpha, \beta, \gamma\}$, so that we also have, trivially, that $\beta \in B_2([a_2] \rightarrow [a_2c_1])$ and $\beta \in B_2([a_2] \rightarrow [a_2c_2])$.



Part of a model of the game of Figure 2*a*

Figure 3

As a condition of rationality, should one rule out situations like the one illustrated in Figure 3? Shouldn't a rational player have introspective access to what she would do in all the relevant hypothetical situations? The answer is negative, since - in general - what a player would do may depend on external circumstances (e.g. the actions of other players) and no amount of introspection can aid the player in forming correct beliefs about these external circumstances. This is illustrated in the following example (inspired by Example 5.1 in [31], p. 325).

Example 1 *Players 1 and 2 are in a dark room, facing a wall which is painted either blue or red (since the light is off, neither player knows what the color of the wall is). Player 1 has to choose between turning the light on (action N) or leaving the light off (action F). If Player 1 chooses F, the game ends; if he chooses N, then Player 2 - who has a box with two buttons, one blue and one red - has to press one of the two buttons. Payoffs are as follows: Player 1 gets a payoff of 1 if the light stays off, and a payoff of 0 if he turns the light on (no matter what Player 2 does). Player 2 gets a payoff of 1 if the*

light is turned on and she presses the button whose color matches the color of the wall, and a payoff of 0 in every other case. Consider a state, call it β, where the wall is red and Player 1 leaves the light off. What would player 2 do if the light were on? The counterfactual selection function must select a state, call it α, where the wall is still red (turning the light on does not change the color of the wall) and in that state Player 2 - if rational - would press the red button. Suppose also that at state β Player 2 is certain that the wall is blue and that Player 1 is going to leave the light off: $\mathcal{B}_2(\beta) = \{\gamma\}$ and at γ the wall is blue and Player 1 leaves the light off. At state γ what would Player 2 do if the light were on? In this case the counterfactual selection function must select a state, call it δ, where the wall is still blue and in that state Player 2 - if rational - would press the blue button. Hence at state β *we have that both of the following propositions are true: (1) if the light were on, Player 2 would press the* red *button and (2) Player 2 believes that, if the light were on, she would press the* blue *button.*

The above remarks highlight the subtleties involved in defining rationality in dynamic games. However, there are further issues. Consider, for example, the perfect-information game of Figure 2a, a model of this game and a state α where Player 1 plays a_2 ($\alpha \in [a_2]$). Is a_2 a rational choice for Player 1? Answering this question requires answering the following two questions:

Q1. What *will* Player 2 do next?

Q2. What *would* Player 2 do if, instead, a_1 had been chosen?

Let us start with Q1. Suppose that at α the play of the game is $a_2c_2d_1$ (that is, $\zeta(\alpha) = a_2c_2d_1$) and that Player 1 has correct beliefs about this:

$\mathcal{B}_1(\alpha) = \{\alpha\}$. If there is "common recognition" of rationality, Player 1 will ask himself how a rational Player 2 will respond to his initial choice of a_2. In order to determine what would be rational for Player 2 to do, we need to examine Player 2's beliefs. Suppose that Player 2 mistakenly believes that Player 1 will play a_1 ($\alpha \in B_2[a_1]$): for example, $\mathcal{B}_2(\alpha) = \{\beta\}$ and $\beta \in [a_1 b_1]$. Furthermore, suppose that $f(\beta, [a_2]) = \gamma$ and $\gamma \in [a_2 c_2 d_2]$. Then at α Player 2 believes that if it were the case that Player 1 played a_2 then the play of the game would be $a_2 c_2 d_2$ ($\alpha \in B_2([a_2] \rightrightarrows [a_2 c_2 d_2])$), in particular, Player 1 would end the game by playing d_2. Since, at state α, Player 1 in fact plays a_2, Player 2 will be surprised: she will be informed that Player 1 played a_2 and that she herself has to choose between c_1 and c_2. What choice she will make depends on her beliefs after she learns that (contrary to her initial expectation) Player 1 played a_2, that is, on her *revised beliefs*. In general, no restrictions can be imposed on Player 2's revised beliefs after a surprise: for example, it seems perfectly plausible to allow Player 2 to become convinced that the play of the game will be $a_2 c_2 d_1$; in particular, that Player 1 will end the game by playing d_1. The models that we are considering do not provide us with the tools to express such a change of mind for Player 2: if one takes as her revised beliefs her initial beliefs about counterfactual statements that have a_2 as an antecedent, then - since $\alpha \in B_2([a_2] \rightrightarrows [a_2 c_2 d_2])$ - one is forced to ruled out the possibility that after learning that Player 1 played a_2 Player 2 will believe that the play of the game will be $a_2 c_2 d_1$. Stalnaker argues that imposing such restrictions is conceptually wrong, since it is based on confounding causal with epistemic counterfactuals:

> "Player 2 has the following initial belief: Player 1 would choose d_2 on his second move [after his initial choice of a_2] if he had a second

> move. This is a causal 'if' – an 'if' used to express 2's opinion about 1's disposition to act in a situation that she believes will not arise. [...] But to ask what Player 2 would believe about Player 1 if she learned that she was wrong about 1's first choice is to ask a completely different question – this 'if' is epistemic; it concerns Player 2's belief revision policies, and not Player 1's disposition to act." ([43], p. 48; with small changes to adapt the quote to the game of Figure 1a.)

Let us now turn to question Q2 and continue the above example, where

$$\alpha \in [a_2c_2d_1] \cap B_1([a_2c_2d_1]) \cap B_2([a_1b_1]) \cap B_1B_2([a_1b_1]). \tag{9}$$

Thus at α Player 1 plays a_2. Is this a rational choice? The answer depends on how Player 2 would respond to the alternative choice of a_1. However, since the rationality of playing a_2 has to be judged relative to Player 1's beliefs, what matters is not what Player 2 would actually do at state α if a_1 were to be played, but what Player 1 believes that Player 2 would do. How should we model such beliefs of Player 1? Again, one possibility is to refer to Player 1's beliefs about counterfactuals with $[a_1]$ as antecedent. If we follow this route, then we restrict the possible beliefs of Player 1; in particular, it cannot be the case that Player 1 believes that if he were to play a_1 then Player 2 would play b_2, that is, we cannot have $\alpha \in B_1([a_1] \rightrightarrows [a_1b_2])$. The reason is as follows. The counterfactual selection function is meant to capture causal relationships between events. As Stalnaker points out, in the counterfactual world where a player makes a choice different from the one that he is actually making, the prior beliefs of the other players must be the same as in the actual world (by changing his choice he cannot cause the prior beliefs of his opponents to change):

> "I know, for example, that it would be irrational to cooperate in a one-shot prisoners' dilemma because I know that in the counterfactual situation in which I cooperate, my payoff is less than it would be if I defected. And while I have the capacity to influence my payoff (negatively) by making this alternative choice, I could not, by making this choice, influence your prior beliefs about what I will do; that is, your prior beliefs will be the same, in the counterfactual situation in which I make the alternative choice, as they are in the actual situation." ([45], p. 178)

We claimed that it canot be the case that $\alpha \in B_1([a_1] \rightrightarrows [a_1b_2])$. To see this, suppose that $\alpha \in B_1([a_1] \rightrightarrows [a_1b_2])$ and fix an arbitrary $\omega \in \mathcal{B}_1(\alpha)$. By (9), since $\alpha \in B_1([a_2])$, $\omega \in [a_2]$; furthermore, if $\delta = f(\omega, [a_1])$ then, since $\alpha \in B_1([a_1] \rightrightarrows [a_1b_2])$, $\delta \in [a_1b_2]$. Since $\omega \in \mathcal{B}_1(\alpha)$ and $\alpha \in B_1B_2([a_1b_1])$, $\omega \in B_2([a_1b_1])$. By the above remark, at δ the initial beliefs of Player 2 must be the same as at ω. Hence $\delta \in B_2([a_1b_1])$. But this, together with $\delta \in [a_1b_2]$, violates (8).[15]

One approach followed in the literature (see, for example, [2, 7, 8, 12, 19, 20, 26, 31, 38] is to do without an "objective" counterfactual selection function f and introduce in its place "subjective" counterfactual functions f_i, one for each player $i \in N$, representing the players' dispositions to revise their beliefs under various hypotheses.[16] This is the topic of the next section.

[15] By definition, $\delta \in B_2([a_1b_1])$ if and only if $\mathcal{B}_2(\delta) \subseteq [a_1b_1]$. Thus, since $[a_1b_1] \subseteq \neg[a_1] \cup [a_1b_1] = [a_1] \rightarrow [a_1b_1]$, $\mathcal{B}_2(\delta) \subseteq [a_1] \rightarrow [a_1b_1]$, that is, $\delta \in B_2([a_1] \rightarrow [a_1b_1])$. Now, (8) requires that, since $\delta \in [a_1b_2]$, $\delta \in B_2([a_1] \rightarrow [a_1b_2])$, yielding a contradiction.

[16] In [26] there is also an objective counterfactual selection function, but it used only to encode the structure of the game in the syntax.

4 Belief revision

We will now consider models of dynamic games defined as tuples $\left\langle N, \Omega, \{\mathcal{B}_i\}_{i\in N}, \{\mathcal{E}_i, f_i\}_{i\in N}, \zeta \right\rangle$ where - as before - $\left\langle N, \Omega, \{\mathcal{B}_i\}_{i\in N} \right\rangle$ is an interactive belief structure, $\zeta : \Omega \to Z$ is a function that associates with every state $\omega \in \Omega$ a terminal history and, for every player $i \in N$, $\mathcal{E}_i \subseteq 2^\Omega \backslash \varnothing$ is a set of events representing potential items of information or admissible hypotheses for player i [17] and, $f_i : \Omega \times \mathcal{E}_i \to 2^\Omega$ is a function such that, $\forall \omega \in \Omega$, $\forall E, F \in \mathcal{E}_i$,

$$\begin{array}{ll} 1. & f_i(\omega, E) \neq \varnothing, \\ 2. & f_i(\omega, E) \subseteq E, \\ 3. & \text{if } \mathcal{B}_i(\omega) \cap E \neq \varnothing \text{ then } f_i(\omega, E) = \mathcal{B}_i(\omega) \cap E, \\ 4. & \text{if } E \subseteq F \text{ and } f_i(\omega, F) \cap E \neq \varnothing \text{ then } f_i(\omega, E) = f_i(\omega, F) \cap E. \end{array} \tag{10}$$

The interpretation of $f_i(\omega, E)$ is the set of states that player i would consider possible under the supposition that (or if informed that) E is true. Condition 1 requires these suppositional beliefs to be consistent. Condition 2 requires that E be indeed considered true. Condition 3 says that if E is compatible with the initial beliefs then the suppositional beliefs coincide with the initial beliefs conditioned on event E.[18] Condition 4 is an extension of

[17] For example, in a perfect-information game one can take $\mathcal{E}_i = \{[h] : h \in H_i\}$, that is, the set of propositions of the form "decision history h of player i is reached" or $\mathcal{E}_i = \{[h] : h \in H\}$, the set of propositions corresponding to all histories (in which case $\mathcal{E}_i = \mathcal{E}_j$ for any two players i and j).

[18] Note that it follows from Conditions 1 and 3 (and seriality of $\mathcal{B}_i$) that, for every $\omega \in \Omega$, $f_i(\omega, \Omega) = \mathcal{B}_i(\omega)$, so that one could simplify the definition of model by dropping the relations $\mathcal{B}_i$ and recovering the initial beliefs from the set $f_i(\omega, \Omega)$. We have chosen not to do so in order to maintain continuity in the exposition.

3: if E implies F and E is compatible not with player i's prior beliefs but with the *posterior* beliefs that she would have if she supposed (or learned) that F were the case (let's called these her posterior F-beliefs), then her beliefs under the supposition (or information) that E must coincide with her posterior F-beliefs conditioned on even E.[19]

Remark 2 *If $\mathcal{E}_i = 2^\Omega \backslash \varnothing$ then Conditions 1-4 in (10) imply that, for every $\omega \in \Omega$, there exists a "plausibility" relation Q_i^ω on Ω which is complete ($\forall \omega_1, \omega_2 \in \Omega$, either $\omega_1 Q_i^\omega \omega_2$ or $\omega_2 Q_i^\omega \omega_1$ or both) and transitive ($\forall \omega_1, \omega_2, \omega_3 \in \Omega$, if $\omega_1 Q_i^\omega \omega_2$ and $\omega_2 Q_i^\omega \omega_3$ then $\omega_1 Q_i^\omega \omega_3$) and such that, for every $E \subseteq \Omega$ with $E \neq \varnothing$, $f_i(\omega, E) = \{x \in E : x Q_i^\omega y, \ \forall y \in E\}$. The interpretation of $\alpha Q_i^\omega \beta$ is that - at state ω and according to player i - state α is at least as plausible as state β. Thus $f_i(\omega, E)$ is the set of most plausible states in E (according to player i at state ω). If $\mathcal{E}_i \neq 2^\Omega \backslash \varnothing$ then Conditions 1-4 in (10) are necessary but not sufficient for the existence of such a plausibility relation. The existence of a plausibility relation that rationalizes the function $f_i(\omega, \cdot) : \mathcal{E}_i \to 2^\Omega$ is necessary and sufficient for the belief revision policy encoded in $f_i(\omega, \cdot)$ to be compatible with the theory of belief revision introduced in [1], known as the AGM theory (see [18]).*

One can associate with each function f_i an operator $\rightrightarrows_i : \mathcal{E}_i \times 2^\Omega \to 2^\Omega$ as follows:

$$E \rightrightarrows_i F = \{\omega \in \Omega : f_i(\omega, E) \subseteq F\}. \tag{11}$$

[19] Although widely accepted, this principle of belief revision is not uncontroversial (see [36, 46]).

Possible interpretations of the event $E \rightrightarrows_i F$ are "according to player i, if E were the case, then F would be true" ([31]) or "if informed that E, player i would believe that F" ([43]) or "under the supposition that E, player i would believe that F" ([2]).[20]

Thus the function f_i can be used to model the full epistemic state of player i: in particular, how player i would revise her prior beliefs if she acquired information that contradicted those beliefs. It is important to note, however, that even with the addition of the functions f_i, the models remain *static* in nature: they represent only the players' beliefs at a fixed point in time (before the game is played), together with their dispositions to revise those beliefs. Thus these models do not represent any actual revisions that are made when new information is actually received.

Condition (8) rules out the possibility that a player may be uncertain about her own choice of action at decision histories of hers that are not ruled out by her initial beliefs. Does a corresponding restriction hold for revised beliefs? That is, suppose that at a state ω player i erroneously believes that her decision history h will not be reached ($\omega \in [h]$ but $\omega \in B_i \neg [h]$); suppose also that a is the action that she will choose at h ($\omega \in [ha]$). Is it necessarily the case that, according to her revised beliefs on the suppositions that h is reached, she believes that she takes action a? That is, is it the case that $[h] \rightrightarrows_i [ha]$? In general, the answer is negative. For example, consider the game of Figure 2*a* and states α, β and γ such that $\alpha \in [a_1 b_1]$, $\mathcal{B}_2(\alpha) = \{\beta\}$,

[20] Equivalently, one can think of $\rightrightarrows_i$ as a conditional belief operator $B_i(\cdot|\cdot)$ with the interpretation of $B_i(F|E)$ as 'player i believes F given information/supposition E'. An alternative notation that can be found in the literature for $B_i(F|E)$ is $B_i^E(F)$ (see, for example, [13]).

$\beta \in [a_2c_1]$, $f_2(\alpha, [a_1]) = \{\gamma\}$ and $\gamma \in [a_1b_2]$. Then we have that at state α Player 2 will in fact take action b_1 (after being surprised by Player 1's choice of a_1) and yet, according to her revised beliefs on the supposition that Player 1 plays a_1, she does not believe that she would take action b_1 (in fact she believes that she would take action b_2): $\alpha \notin [a_1] \rightrightarrows_i [a_1b_1]$. In order to rule this out we need to impose the following strengthening of (8):[21]

$$\forall h \in H_i,\ \forall a \in A(h), \quad [ha] \subseteq ([h] \rightrightarrows_i [ha])\,. \tag{12}$$

Can (12) be considered a necessary component of a definition of rationality? Perhaps so if the revised beliefs are interpreted as the actual beliefs of player i when she is actually informed (to her surprise) that her decision history h has been reached. In that case it seems reasonable to assume that - as the player makes up her mind about what to do - she forms correct beliefs about what she is going to do. However, we stressed above that the above models are static models: they represent the initial beliefs and disposition to revise those beliefs at the beginning of the game. Given this interpretation of the revised beliefs as hypothetical beliefs conditional on various suppositions, it seems that violations of (12) might be perfectly rational. To

[21] (12) is implied by (8) whenever player i's initial beliefs do not rule out h. That is, if $\omega \in \neg B_i \neg [h]$ (equivalently, $\mathcal{B}_i(\omega) \cap [h] \neq \varnothing$) then, for every $a \in A(h)$,

$$\text{if } \omega \in [ha] \text{ then } \omega \in ([h] \rightrightarrows_i [ha])\,. \quad \text{(F1)}$$

In fact, if $\mathcal{B}_i(\omega) \cap [h] \neq \varnothing$ then by Condition 3 of (10),

$$f_i(\omega, [h]) = \mathcal{B}_i(\omega) \cap [h]. \quad \text{(F2)}$$

Let $a \in A(h)$ be such that $\omega \in [ha]$. Then, by (8), $\omega \in B_i([h] \rightarrow [ha])$, that is, $\mathcal{B}_i(\omega) \subseteq \neg[h] \cup [ha]$. Thus $\mathcal{B}_i(\omega) \cap [h] \subseteq (\neg[h] \cap [h]) \cup ([ha] \cap [h]) = \varnothing \cup [ha] = [ha]$ (since $[ha] \subseteq [h]$) and therefore, by (F2), $f_i(\omega, [h]) \subseteq [ha]$, that is, $\omega \in [h] \rightrightarrows_i [ha]$.

illustrate this point, consider the above example with the following modification: $f_2(\alpha, [a_1]) = \{\alpha, \gamma\}$. It is possible that if Player 1 plays a_1, Player 2 is indifferent between playing b_1 or b_2 (she gets the same payoff). Thus she can coherently form the belief that if - contrary to what she expects - Player 1 were to play a_1, then she might end up choosing either b_1 or b_2: $\alpha \in [a_1] \rightrightarrows_i ([a_1b_1] \cup [a_1b_2])$. Of course, when she will actually be faced with the choice between b_1 and b_2 she will have to break her indifference and pick one action (perhaps by tossing a coin): in this case she will pick b_1 (perhaps because the outcome of the coin toss will be Heads: something she will know then but cannot know at the beginning).

How can rationality of choice be captured in the models that we are considering? Various definitions of rationality have been suggested in the literature, most notably *material rationality* and *substantive rationality* ([4, 5]). The former notion is weaker in that a player can be found to be irrational only at decision histories of hers that are actually reached. The latter notion, on the other hand, is more stringent since a player can be judged to be irrational at a decision history h of hers even if she knows that h will not be reached. We will focus on the weaker notion. We want to define a player's rationality as a proposition, that is, an event. Let $u_i : Z \to \mathbb{R}$ be player i's ordinal utility function (representing her preferences over the set of terminal histories Z) and define $\pi_i : \Omega \to \mathbb{R}$ by $\pi_i(\omega) = u_i(\zeta(\omega))$. For every $x \in \mathbb{R}$, let $[\pi_i \leq x]$ be the event that player i's payoff is not greater than x, that is, $[\pi_i \leq x] = \{\omega \in \Omega : \pi_i(\omega) \leq x\}$ and, similarly, let $[\pi_i > x] = \{\omega \in \Omega : \pi_i(\omega) > x\}$. Then we say that player i is rational at a state if, for every decision history of hers that is actually reached at that state and for every

real number x, it is not the case that she believes that her payoff is not greater than x and it would be greater than x if she were to take an action different from the one that she is actually taking (at that history in that state). Formally this can be stated as follows (recall that H_i denotes the set of decision histories of player i and $A(h)$ the set of actions available at h):

$$\begin{array}{l} \text{Player } i \text{ is rational at } \omega \in \Omega \text{ if, } \forall h \in H_i, \forall a \in A(h) \\ \text{if } ha \text{ is a prefix of } \zeta(\omega) \text{ then, } \forall b \in A(h),\ \forall x \in \mathbb{R}, \\ \left([ha] \rightrightarrows_i [\pi_i \leq x]\right) \rightarrow \neg \left([hb] \rightrightarrows_i [\pi_i > x]\right). \end{array} \tag{13}$$

Note that, in general, we cannot replace the antecedent $[ha] \rightrightarrows_i [\pi_i \leq x]$ with $B_i([ha] \rightarrow [\pi_i \leq x])$, because at state ω player i might initially believe that h will not be reached, in which case it would be trivially true that $\omega \in B_i([ha] \rightarrow [\pi_i \leq x])$; however, if decision history h is actually reached at ω then player i will be surprised and will have to revise her beliefs, given the information that h has been reached. Thus, in general, her rationality is judged on the basis of her *revised* beliefs. Note, however, that if $\omega \in \neg B_i \neg [h]$, that is, if at ω she does not rule out the possibility that h will be reached and $a \in A(h)$ is the action that she actually takes at ω ($\omega \in [ha]$), then, for

every event F, $\omega \in B_i([ha] \rightarrow F)$ if and only if $\omega \in ([ha] \rightrightarrows_i F)$.[22] Note also that, according to (13), a player is trivially rational at any state at which she does not take any actions.

Does initial common belief that all the players are materially rational imply backward induction in perfect-information games? The answer is negative; in fact, common belief of material rationality does not even imply a Nash equilibrium outcome. To see this, consider the perfect-information game shown in Figure 4*a* and the model of it shown in Figure 4*b*.[23] First of all, note that the common belief relation $\mathcal{B}^*$ is obtained by adding to $\mathcal{B}_2$ the pair (β, β); thus, in particular, $\mathcal{B}^*(\beta) = \{\beta, \gamma\}$. We want to show that both players are materially rational at both states β and γ, so that at state β it is common belief that both players are materially rational, despite that fact that the play of the game at β is $a_1a_2d_3$, which is not sustained by any

[22] Proof. Suppose that $\omega \in [ha] \cap \neg B_i \neg [h]$. As shown in Footnote 21 (see (F2)),

$$\mathcal{B}_i(\omega) \cap [h] = f_i(\omega, [h]). \quad \text{(G1)}$$

Since $[ha] \subseteq [h]$,

$$\mathcal{B}_i(\omega) \cap [h] \cap [ha] = \mathcal{B}_i(\omega) \cap [ha] \quad \text{(G2)}$$

As shown in Footnote 21, $f_i(\omega, [h]) \subseteq [ha]$ and, by Condition 1 of (10), $f_i(\omega, [h]) \neq \varnothing$. Thus $f_i(\omega, [h]) \cap [ha] = f_i(\omega, [h]) \neq \varnothing$.Hence, by Condition 4 of (10),

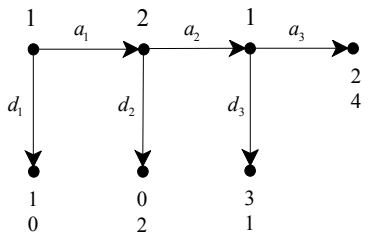
$$f_i(\omega, [h]) \cap [ha] = f_i(\omega, [ha]). \quad \text{(G3)}$$

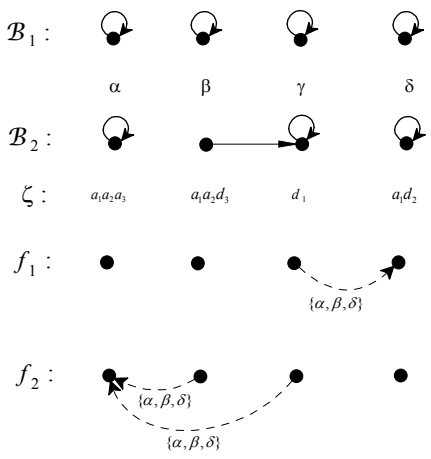
By intersecting both sides of (G1) with $[ha]$ and using (G2) and (G3) we get that $\mathcal{B}_i(\omega) \cap [ha] = f_i(\omega, [ha])$. ■

[23] In Figure 4*a*, for every terminal history, the top number associated with it is Player 1's utility and the bottom number is Player 2's utility. In Figure 4*b* we have only represented parts of the functions f_1 and f_2.

Similar examples can be found in [13, 26, 37, 43].

Nash equilibrium. Clearly, Player 1 is materially rational at state β (since he obtains his largest possible payoff); he is also rational at state γ because he knows that he plays d_1, obtaining a payoff of 1, and believes that if he were to play a_1 Player 2 would respond with d_2 and give him a payoff of zero: this belief is encoded in $f_1(\gamma, [a_1]) = \{\delta\}$, where $[a_1] = \{\alpha, \beta, \delta\}$. Player 2 is trivially materially rational at state γ since she does not take any actions there. Now consider state β. Player 2 initially erroneously believes that Player 1 will end the game by playing d_1; however, Player 1 is in fact playing a_1 and thus Player 2 will be surprised. Her initial disposition to revise her beliefs on the supposition that Player 1 plays a_1 is such that she would believe that she herself would play a_2 and Player 1 would follow with a_3, thus giving her the largest possible payoff. Hence she is rational at state β, according to (13).





(*a*) A perfect-information game (*a*) A model of the game

Figure 4

Thus, in order to obtain the backward-induction solution, one needs to go beyond common initial belief of material rationality. Proposals in the

literature include the notions of epistemic independence ([43]), strong belief ([10]), stable belief ([7]), substantive rationality ([4, 33]). For an overview of this literature the reader is referred to [23, 35].

In the models considered above, strategies do not play any role: states are described in terms of the players' actual behavior along a play of the game. One could view a player's strategy as her (revised) beliefs about what she would under the supposition that each of her decision histories is reached. However, the models considered so far do not guarantee that a player's revised beliefs select a unique action at each of her decision histories. For example, consider the game of Figure 2*a* and states α, β and γ such that $\alpha \in [a_2c_1]$, $\mathcal{B}_2(\alpha) = \{\alpha\}$, $\beta \in [a_1b_1]$, $\gamma \in [a_1b_2]$ and $f_2(\alpha, [a_1]) = \{\beta, \gamma\}$. Then at state α Player 2 knows that she will take action c_1 and, according to her revised beliefs on the supposition that Player 1 plays a_1, she is uncertain as to whether she would respond to a_1 by playing b_1 or b_2 (perhaps she is indifferent between b_1 and b_2, because she would get the same payoff in either case). One could rule this possibility out by imposing the following restriction;

$$\begin{array}{l} \forall h \in H_i, \forall a, b \in A(h), \forall \omega, \omega', \omega'' \in \Omega, \text{ if } \omega', \omega'' \in f_i(\omega, [h]) \\ \text{and } ha \text{ is a prefix of } \zeta(\omega') \text{ and } hb \text{ is a prefix of } \zeta(\omega'') \text{ then } a = b. \end{array} \tag{14}$$

If (14) is imposed then one can associate with every state a unique strategy for every player. As Samet points out ([38], p. 232) in this setup strategies are cognitive constructs rather than objective counterfactuals about what a player would actually do at each of her decision histories.

5 Conclusion

Roughly speaking, a player's choice is rational if, according to what the player believes, there is no other choice which is better for her. Thus, in order to be able to assess the rationality of a player one needs to be able to represent both the player's choices and her beliefs. The notion of model of a game does precisely this. We have discussed a number of conceptual issues that arise in attempting to represent not only the actual beliefs but also the counterfactual or hypothetical beliefs of the players. These issues highlight the complexity of defining the notion of rationality in dynamic games and of specifying an appropriate interpretation of the hypothesis that there is "common recognition" of rationality.

The models of dynamic games considered above are not the only possibility. Instead of modeling the epistemic states of the players in terms of their prior beliefs and prior dispositions to revise those beliefs, one could model the actual belief revision taking place during the play of the game (for example, by using the structures introduced in [21]). Alternatively, one could model (conditional) beliefs using the notion of prediction in branching-time frames introduced in [15, 16]. Because of space limitations we do not pursue these possibilities here.

The focus of this chapter has been on the issue of modelling the notion of rationality and "common recognition" of rationality in dynamic games. Alternatively one can use the AGM theory of belief revision to provide foundations for refinements of Nash equilibrium in dynamic games. This is done in [19, 20] where a general notion of perfect Bayesian equilibrium is proposed for general dynamic games (thus allowing for imperfect information). Perfect

Bayesian equilibria constitute a refinement of subgame-perfect equilibria and are a superset of sequential equilibria.

References

[1] Alchourrón, Carlos, Peter Gärdenfors and David Makinson, On the logic of theory change: partial meet contraction and revision functions, *The Journal of Symbolic Logic*, 1985, 50: 510-530.

[2] Arló-Costa, Horacio and Cristina Bicchieri, Knowing and supposing in games of perfect information, *Studia Logica,* 2007, 86: 353-373.

[3] Aumann, Robert, What is game theory trying to accomplish?, in Kenneth Arrow and Seppo Hinkapohja (editors), *Frontiers in economics,* Basil Blackwell, Oxford, 1985, 28-76.

[4] Aumann, Robert, Backward induction and common knowledge of rationality, *Games and Economic Behavior*, 1995, 8: 6–19.

[5] Aumann, Robert, On the centipede game, *Games and Economic Behavior*, 1998, 23: 97-105.

[6] Aumann, Robert and Adam Brandenburger, Epistemic conditions for Nash equilibrium, *Econometrica*, 1995, 63: 1161-1180.

[7] Baltag, Alexandru, Sonja Smets and Jonathan Zvesper, Keep 'hoping' for rationality: a solution to the backward induction paradox, *Synthese*, 2009, 169: 301-333.

[8] Battigalli, Pierpaolo, Alfredo Di Tillio and Dov Samet, Strategies and interactive beliefs in dynamic games, Technical Report IGIER WP 375, Bocconi University, 2011.

[9] Battigalli, Pierpaolo and Giacomo Bonanno, Recent results on belief, knowledge and the epistemic foundations of game theory, *Research in Economics*, 1999, 53: 149-225.

[10] Battigalli, Pierpaolo and Marciano Siniscalchi, Strong belief and forward induction reasoning, *Journal of Economic Theory*, 2002, 106: 356-391.

[11] Ben Porath, Elchanan, Rationality, Nash equilibrium, and backwards induction in perfect information games, *Review of Economic Studies*, 1997, 64: 23–46.

[12] Board, Oliver, Belief revision and rationalizability, in Itzhak Gilboa (editor), *Theoretical aspects of rationality and knowledge (TARK VII)*, 1998, Morgan Kaufman, San Francisco, 201-213.

[13] Board, Oliver, Dynamic interactive epistemology, *Games and Economic Behavior*, 2004, 49: 49-80.

[14] Board, Oliver, The equivalence of Bayes and causal rationality in games, *Theory and Decision*, 2006, 61:1-19.

[15] Bonanno, Giacomo, Prediction in branching time logic, *Mathematical Logic Quarterly*, 2001, 47: 239-247.

[16] Bonanno, Giacomo, Revising predictions, in Johan van Benthem (editor), T*heoretical aspects of rationality and knowledge (TARK 2001)*, 2001, Morgan Kaufman, San Francisco, 273-286.

[17] Bonanno, Giacomo, A syntactic approach to rationality in games with ordinal payoffs. In: Giacomo. Bonanno, Wiebe van der Hoek and Michael Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory*, Texts in Logic and Games, Amsterdam University Press, 2008, 59-86.

[18] Bonanno, Giacomo, Rational choice and AGM belief revision, *Artificial Intelligence*, 2009, 173: 1194-1203.

[19] Bonanno, Giacomo, *AGM* belief revision in dynamic games, in: Krzysztof R. Apt (editor), *Proceedings of the 13th conference on theoretical aspects of rationality and knowledge (TARK XIII)*, ACM, New York, 2011, 37-45 (doi: 10.1145/2000378.2000383)

[20] Bonanno, Giacomo, *AGM*-consistency and perfect Bayesian equilibrium. Part I: definition and properties, *International Journal of Game Theory*, 2011 (doi: 10.1007/s00182-011-0296-4).

[21] Bonanno, Giacomo, Belief change in branching time: AGM-consistency and iterated revision, *Journal of Philosophical Logic*, 2011, doi: 10.1007/s10992-011-9202-6.

[22] Bonanno, Giacomo and Klaus Nehring, Assessing the truth axiom under incomplete information, *Mathematical Social Sciences,* 1998, 36: 3-29.

[23] Brandenburger, Adam, The power of paradox: some recent developments in interactive epistemology, *International Journal of Game Theory*, 2007, 35: 465–492.

[24] Camerer, Colin, *Behavioral game theory: experiments in strategic interaction*, Princeton University Press, Princeton, 2003.

[25] Clausing, Thorsten, Doxastic conditions for backward induction, *Theory and Decision,* 2003, 54: 315-336.

[26] Clausing, Thorsten, Belief revision in games of perfect information, *Economics and Philosophy, 2004,* 20: 89-115.

[27] de Bruin, Boudewijn, *Explaining games: the epistemic programme in game theory*, Springer (Synthese Library), 2010.

[28] Dekel, Eddie and Faruk Gul, Rationality and knowledge in game theory, in: David Kreps and Kenneth Wallis (editors), *Advances in Economics and Econometrics*, Cambridge University Press, Cambridge, 1997, 87-172.

[29] Feinberg, Yossi, Subjective reasoning - dynamic games, *Games and Economic Behavior*, 2005, 52: 54-93.

[30] Gerstung, Moritz, Hani Nakhoul and Niko Beerenwinkel, Evolutionary games with affine fitness functions: applications to cancer, *Dynamic Games and Applications*, 2011, 1: 370-385.

[31] Halpern Joseph, Hypothetical knowledge and counterfactual reasoning, *International Journal of Game Theory,* 1999, 28: 315-330.

[32] Halpern Joseph, Set-theoretic completeness for epistemic and conditional logic, *Annals of Mathematics and Artificial Intelligence*, 1999, 26: 1–27.

[33] Halpern Joseph, Substantive rationality and backward induction, *Games and Economic Behavior*, 2001, 37: 425–435.

[34] Lewis, David, *Counterfactuals,* Harvard University Press, Cambridge, 1973.

[35] Perea, Andrés, Epistemic foundations for backward induction: an overview, in Johan van Benthem, Dov Gabbay and Benedikt Löwe (editors), *Interactive logic. Proceedings of the 7th Augustus de Morgan Workshop,* Texts in Logic and Games 1, Amsterdam University Press, 2007, 159-193.

[36] Rabinowicz, Wlodek, Stable revision, or is preservation worth preserving?, in André Fuhrmann and Hans Rott (editors), *Logic, action and information: essays on logic in philosophy and artificial intelligence,* de Gruyter, Berlin, 1996, 101-128.

[37] Rabinowicz, Wlodek, Backward induction in games: on an attempt at logical reconstruction, in Wlodek Rabinowicz, (Editor), *Value and choice: some common themes in decision theory and moral philosophy,* Lund Philosophy Reports, 2000, 243-256.

[38] Samet, Dov, Hypothetical knowledge and games with perfect information, *Games and Economic Behavior*, 1996, 17: 230–251.

[39] Shoham, Yoav and Kevin Leyton-Brown, *Multiagent systems: algorithmic, game-theoretic, and logical foundations*, Cambridge University Press, Cambridge, 2008.

[40] Smith, John Maynard, *Evolution and the theory of games,* Cambridge University Press, Cambridge, 1982.

[41] Stalnaker, Robert, A theory of conditionals, in N. Rescher (editor), *Studies in logical theory*, Blackwell, 1968, 98-112.

[42] Stalnaker, Robert, Knowledge, belief and counterfactual reasoning in games, *Economics and Philosophy,* 1996, 12: 133-163.

[43] Stalnaker, Robert, Belief revision in games: forward and backward induction, *Mathematical Social Sciences*, 1998, 36: 31–56.

[44] Stalnaker, Robert, Extensive and strategic forms: games and models for games, *Research in Economics,* 1999, 53: 293 -319.

[45] Stalnaker, Robert, On logics of knowledge and belief, *Philosophical Studies,* 2006, 128: 169-199.

[46] Stalnaker, Robert, Iterated belief revision, *Erkenntnis,* 2009, 70: 189-209.

[47] von Neumann, John and Oscar Morgenstern, *Theory of games and economic behavior*. Princeton University Press, 1944.

[48] Zambrano, Eduardo, Counterfactual reasoning and common knowledge of rationality in normal form games, *Topics in Theoretical Economics,* 2004, 4: Article 8.